Ahrii Kim[1]

[1]Independent Researcher

July 19, 2025

IR_MULTI-AGENTMT at WMT25 Translation Task: A Summary

Ahrii Kim
Independent Researcher / Seoul, Korea
ahriikim@gmail.com

Abstract

We introduce our model, referred to as MULTI-AGENTMT, for participation in the WMT 25 Translation Task. This model operationalizes the notion of an AI Agent by employing a multi-agent workflow known as Prompt Chaining (Briva-Iglesias, 2025) alongside the automatic MQM error annotation framework designated as RUBRIC-MQM (Kim, 2025). Our primary submission is developed through the Translate-Postedit-Proofread paradigm, whereby each stage incrementally enhances the translation output. Our experimental findings indicate the feasibility of implementing a semi-autonomous improvement process in Machine Translation within this framework, yielding superior outcomes with a smaller model at reduced cost.

1 Introduction

An AI agent is a computational system that operates autonomously, guided by environmental observations, and frequently incorporates adaptive learning abilities (Russell and Norvig, 2010). Large Language Models (LLM), essential in advancing AI agents, boast enhanced reasoning, contextual comprehension, and adaptable workflows across diverse tasks, notably including Machine Translation (MT) (Briva-Iglesias, 2025). Briva-Iglesias (2025) examined a multi-agent system comprising four agents: Translator, Reviewers of Fluency and Adequacy, and Editor, illustrating its promising potential for future advancements. Drawing inspiration from this innovative and forward-thinking approach, we are participating in this year's MT track with a multi-agent workflow. Our objective is to leverage smaller models to outperform larger models while reducing computational expenses.

2 Participating Task

The primary objective of the competition is to evaluate performance across a diverse array of languages, domains, genres, and modalities. We are participating in a **multilingual subtask** comprising translations of 30 languages, utilizing Czech, English, and Japanese as the source languages, serving as an **unconstrained system**. The source data comprises 29,957 segments, equal to 102,060 paragraphs. Our approach involves obtaining translation inference on a segment-by-segment basis; however, in instances where paragraph boundaries are disregarded, we separate them into paragraphs to obtain results independently.

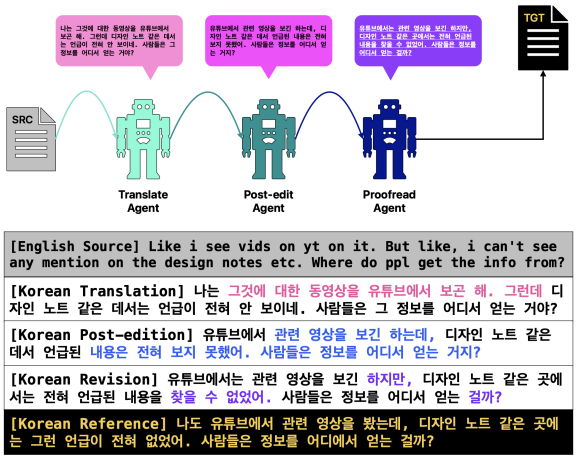

Figure 1: The operational mechanism of MULTI-AGENTMT within the Translate-Postedit-Proofread framework. Each agent sequentially accepts the output of its predecessor as input, refining the translation iteratively. As an illustration, an English-to-Korean translation instance from WMT 2025 dataset, accompanied by a reference translation of our own formulation, is presented.

3 MULTI-AGENTMT

3.1 Design

AI agents facilitate dynamic workflows through configurable architectures. We utilize the concept of Prompt Chaining, forming an orderly and consecutive process where each step's result feeds into the next, promoting systematic reasoning and on-

Method	Translate Tokens	Post-edit Tokens	Proofread Tokens	Cost ($)
Translate	10,728,181	–	–	2.00
Post-edit	–	58,548,177	–	10.89
Proofread	–	–	15,874,680	2.95
TP (Translate + Post-edit)	10,728,181	58,548,177	–	12.89
TPP (Translate + Post-edit + Proofread)	10,728,181	58,548,177	15,874,680	15.84
GPT-4o	10,728,181⋆	–	–	33.26
GPT-o1	10,728,181⋆	–	–	605.07

Table 1: Comparison of token usage and cost among our two workflows (TP, TPP) and bigger GPT models. On the assumption that the larger model employs an equivalent number of tokens as indicated in ⋆, our models (TP and TPP) consume more tokens, but exhibit enhanced cost-efficiency compared to their counterparts while delivering superior translation quality.

going refinement (Briva-Iglesias, 2025). While theoretically, the iterative refinement could enhance translation quality, cost considerations have led us to adopt a unidirectional configuration. Consequently, we examine two types of multi-agent workflows: Translate-Postedit (**TP Workflow**) and Translate-Postedit-Proofread (**TPP Workflow**) as depicted in Figure 1. Each scenario is submitted for the competition.

3.2 Translate Agent

The model is tasked with translating the provided source text according to the officially provided prompt from the organizer. For reasons of cost-effectiveness, it is feasible to obtain translations from more economical platforms such as Google Translate or DeepL. Nonetheless, our preliminary research indicates that a high-quality initial translation produces superior outcomes.

3.3 Post-edit Agent

The model is tasked with post-editing the provided translation with reference to the source text. RUBRIC-MQM (Kim, 2025) serves as the foundational framework, identifying MQM-style error categories, severities, and error spans akin to its counterpart GEMBA-MQM (Kocmi and Federmann, 2023). Nonetheless, Rubric-MQM alleviates the biases linked to MAJOR and MISTRANSLATION labels, enhances the determination of NO-ERROR cases, and augments the precision of error detection.

Three key aspects are revised from the original Rubric-MQM. Firstly, in order to transform the model into a post-editor, we instruct the model to propose translations for each identified error span. Secondly, we adjust the severity scale from 100 to 4. This modification is implemented because the severity level is not our primary focus. However, Kim (2025) has emphasized that the rubric scheme is vital for the model's efficacy. Consequently, the comprehensive rubric is simplified in the prompt. Thirdly, while preserving the original in-context learning examples, we alter one instance from an English-to-German translation direction to a Japanese-to-Korean one, thereby enabling a multilingual configuration in the X-Y direction. This change enhances the model's ability to recognize cases with the NO-ERROR label, preventing it from incorrectly suggesting error-free phrases as an answer. And then, the suggested spans are manually interchanged, with the resultant sentence being considered the final output.

3.4 Proofread Agent

The model is involved in the examination and refinement of the provided translation, and we employ a Chain-of-Thought prompting strategy (Wei et al., 2022). Initially, the model is guided to identify errors, followed by a request to propose five alternative phrases that prioritize fluency of its own and alignment with the source text. Ultimately, the most suitable translation is selected as the final version. It is expected that this methodology will improve any awkward expressions resulting from manual alterations in the preliminary phase.

4 Model Details

We employ prompt engineering for all agents, adopting `GPT-4o-mini-2024-07-18` as the primary baseline, with a temperature parameter set at 1 and a maximum token length of 1024. While this configuration is not optimal regarding reproducibility, our iterative pilot studies have demonstrated that such settings enable the model to explore a greater range of errors and effect modifications, contributing to improved outcomes. Future efforts

will focus on establishing a more stable environment in this context. Additionally, our model is constructed using the AI-agent-customized library on LangGraph[1], incorporating the same GPT model, and we compare the outcomes.

The comprehensive number of tokens and associated costs incurred for this task are detailed in the Table 1, and the pricing information is presented in Table 2. In this study, the average ratio between input and output is represented by $0.08 : 0.92$. Within our multi-agent framework, a significant portion of the tokens is allocated for post-editing ($10.89 out of $15.84). Notwithstanding, it is assumed to be more cost-effective compared to acquiring translations from larger models such as `GPT-4o` and `GPT-o1`, by at least a factor of two, while offering superior quality.

	`4o-mini`	`GPT-4o`	`GPT-o1`
Input	$0.15	$2.50	$15.00
Output	$0.60	$10.00	$60.00

Table 2: Pricing details per 1M input/output tokens for various GPT models (OpenAI).

5 Conclusion

Our investigation on WMT24++ (Deutsch et al., 2025) indicates that the Post-edit Agent is integral to the translation process, underscoring the significance of delineating errors within a structured framework to improve translation quality. Furthermore, although iterative processes incur substantial computational expenses concerning both financial and temporal resources, our empirical findings reveal that the outcome surpasses that of big models with reduced costs and time commitments. These findings will be elaborated upon in the main system paper.

References

Vincent Briva-Iglesias. 2025. Are ai agents the new machine translation frontier? challenges and opportunities of single-and multi-agent systems for multilingual digital communication. *arXiv preprint arXiv:2504.12891*.

Daniel Deutsch, Eleni Briakou, Isaac Caswell, Max Finkelstein, Roni Galor, and 1 others. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Ahrii Kim. 2025. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In *ACL 2025 Industry Track*. Associations for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Stuart Russell and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach*, 3rd edition. Prentice Hall, Upper Saddle River, NJ.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

[1] `http://github.com/langchain-ai/langgraph`